高等学校大数据技术与应用规划教材

数据可视化与分析基础

SHUJU KESHIHUA YU FENXI JICHU

张丹珏　主　编
郑　俊　副主编
顾顺德　主　审

中国铁道出版社有限公司
CHINA RAILWAY PUBLISHING HOUSE CO., LTD.

高等学校大数据技术与应用规划教材

数据可视化与分析基础

SHUJU KESHIHUA YU FENXI JICHU

张丹珏　主　编
郑　俊　副主编
顾顺德　主　审

中国铁道出版社有限公司
CHINA RAILWAY PUBLISHING HOUSE CO., LTD.

内 容 简 介

本书以循序渐进的方式，由浅入深地讲述了数据分析的整个过程。全书共分6章，主要内容包括：数据分析概述、数据可视化初步、数据可视化进阶、数据挖掘基础、数据分析报告和综合案例。每章内附有实用性范例供读者练习，巩固所学知识。

本书在讲解数据可视化的基础性原理的同时，融入真实案例分析，具有较强的实用性，帮助读者举一反三，真正学会大数据可视化和数据挖掘的工具软件，能运用大数据思维解决学习和工作中的实际问题。

本书适合作为高等学校非计算机相关专业大数据可视化、数据媒体设计等课程的教材，也可作为对数据分析感兴趣读者的参考用书。

图书在版编目（CIP）数据

数据可视化与分析基础 / 张丹珏主编 . —北京：中国铁道出版社有限公司，2019.8

高等学校大数据技术与应用规划教材

ISBN 978-7-113-25989-1

Ⅰ. ①数… Ⅱ. ①张… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字(2019)第 174453 号

书　　名：数据可视化与分析基础
作　　者：张丹珏

策　　划：曹莉群　　**编辑部电话：**（010）63589185 转 2007
责任编辑：陆慧萍　卢　笛
封面设计：刘　颖
责任校对：张玉华
责任印制：郭向伟

出版发行：中国铁道出版社有限公司（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.tdpress.com/51eds/
印　　刷：北京柏力行彩印有限公司
版　　次：2019 年 8 月第 1 版　2019 年 8 月第 1 次印刷
开　　本：787 mm×1 092 mm　1/16　**印张：**12.5　**字数：**294 千
书　　号：ISBN 978-7-113-25989-1
定　　价：39.00 元

PREFACE

前言

大数据技术经历了多年的发展，已经在金融、电信、教育、医药等领域得到了较多也较为成功的应用，这使人们看到了该技术所带来的挑战与改革，而 IT 技术的高速发展使得该技术趋于大众化，使得越来越多的人能够参与其中，分享该技术带来的乐趣。

本书系统地介绍了数据分析、数据可视化与数据挖掘的概念和方法，在内容编排上侧重于应用，用案例将知识点进行串联，以期达到提高读者的学习兴趣、增强实践动手能力的目的。

本书对于初次接触数据分析的读者会有很大帮助，书中对数据分析的每一步操作都有详尽的说明，且选用的软件都是相关工具软件，无须编程基础即可完成整个分析过程，使读者能够脱离枯燥的代码环境，专注于数据本身，为数据分析带来全新的思路和视角。书中涉及的数据均来自于网络，仅供学习研究使用。

本书由张丹珏任主编，郑俊任副主编，施庆、赵任颖、程五生、盛家骏、翁少逸和蒋雨蔚参与编写。全书由顾顺德主审。具体分工如下：第 1 章由施庆编写，第 2 章的 2.1~2.5 由程五生编写，第 2 章的 2.6 和第 3 章由郑俊编写，第 4~6 章和附录 A 由张丹珏编写，附录 B 由盛家骏、翁少逸和蒋雨蔚编写，附录 C 由赵任颖编写。

在本书的编写过程中，得到了许多老师的大力支持和热情帮助，中国铁道出版社有限公司对本书的出版给予了大力支持，在此表示衷心的感谢！

由于时间仓促，编者水平有限，书中难免存在疏漏或不足之处，恳请读者批评指正，以便及时修改和完善。

编　者

2019 年 6 月

CONTENTS

目 录

第1章
数据分析概述

在当今飞速发展的数字化社会，数据量呈现井喷式增长，如何从这些数据中提取有效信息显得尤为重要和迫切。一个专业的数据分析师，除了需要掌握各项操作技能，了解各种数据分析工具，更重要的是具备数据分析的思维逻辑。

本章将着重介绍数据分析领域的相关概念、工具及方法，帮助读者了解大数据、数据可视化、数据挖掘、数据分析的步骤、方法和分析法则，为后续的学习打下扎实的理论基础。

1.1 大数据简介

大数据（Big Data）又称巨量资料，是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资源。

大数据具有以下 5V 特征：

（1）Volume（大量）：指的是巨大的数据量，包括采集、存储及计算过程中的数据。大数据的起始计算单位一般是 PB、EB 或 ZB。

其中，数据量的单位换算如下：

1 GB（GigaByte、吉字节）= 1 024 MB；

1 TB（TrillionByte、太字节）= 1 024 GB；

1 PB（PetaByte、拍字节）= 1 024 TB；

1 EB（ExaByte、艾字节）= 1 024 PB；

1 ZB（ZettaByte、泽字节）= 1 024 EB。

（2）Velocity（高速）：指的是数据增长速度快，处理速度也快，时效性要求高。

（3）Variety（多样）：指的是种类和数据来源多样化，包括结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等，多类型的数据对数据的处理能力提出了更高的要求。

（4）Value（价值）：指的是数据价值密度相对较低。随着互联网以及物联网的广泛应用，

信息感知无处不在，而价值密度的高低与数据总量的大小成反比，因此，如何通过强大的机器算法迅速地完成数据的价值“提纯”是目前大数据背景下亟待解决的难题。

（5）Veracity（真实性）：指的是数据的准确性和可信赖度，即数据的质量。

1.2 数据可视化

1.2.1 数据可视化概述

数据可视化旨在借助图形化手段，清晰有效地传达数据中蕴含的信息，其本质是将复杂的数据用视觉展示的方式增强用户对数据的理解，以准确、形象、快速的传达方式凸显数据的含义。数据可视化综合应用计算机科学、图形学、可视化设计、心理学等多个领域的知识，运用符合人类视觉系统的方式为用户提供简洁、直观、形象、有趣、易于理解的数据展示，从而帮助用户了解数据，应用数据。

值得一提的是：数据可视化是一个处于不断演变中的概念，其边界在不断地扩大中，涵盖的范围也变得越来越广。

1.2.2 在线可视化工具

常见的在线可视化工具有以下几种：

1. ECharts

ECharts（网址 https://echarts.baidu.com/）是一个免费的、功能强大的、可视化的库。它可以流畅地运行在 PC 和移动设备上，兼容当前绝大部分浏览器（如 IE 8/9/10/11、Chrome、Firefox、Safari 等），底层依赖轻量级的 Canvas 类库 ZRender，提供直观、生动、可交互、可高度个性化定制的数据可视化图表。简单地说，ECharts 就是一个帮助数据可视化的库。

2. GAPMINDER

GAPMINDER（网址 https://www.gapminder.org/）是位于瑞典斯德哥尔摩的一个非营利机构，他们收集了大量的国际统计数据，用非常简单形象而极富动感的方式进行展示，既可在线播放，又可下载（每次联网时会自动下载更新数据），免费使用。

3. D3

D3（网址 http://d3js.org/）的全称是 Data-Driven Documents，顾名思义是一个被数据驱动的文档，它是一个 JavaScript 函数库，主要用于数据可视化的展现。

4. RAWGraphs

RAWGraphs（网址 https://rawgraphs.io/）号称“电子表格和矢量图形之间的缺失链接”，它建立在 D3.js 之上，界面设计直观，开源免费，不需要任何注册。它有 21 种图表类型的库供选择，所有的处理均在浏览器中完成。此外，RAWGraphs 是高度可定制和可扩展的，甚至可以接受新的自定义布局。

5. Datawrapper

Datawrapper（网址 https://www.datawrapper.de/）是一个用于制作交互式图表的在线数据可视

化工具。通过从 CSV 文件上传数据或直接将其粘贴到字段中，Datawrapper 将生成相关的可视化文件，非常容易使用和生产有效的图形。

6. Tableau Online

Tableau Online（网址 https://www.tableau.com/zh-cn/products/online）是目前较为流行的可视化工具，它支持各种图表、图形、地图和其他图形，是一个完全免费的工具，用户用它制作的图表可以很容易地嵌入任何网页中，无须离开浏览器，即可连接到数据源，也可以使用 Web 制作功能新建工作簿和可视化。此外，Tableau 还有可供下载的付费版本。

7. Plotly

Plotly（网址 https://plot.ly/）是一个开源的 Python 库，可以完成基于 Web 的数据分析和绘图。使用 Plotly 输出的结果是一个使用 Plotly.js 绘制而成的交互网页，同样支持生成静态图表，如 pdf、png 等。

8. Visualize Free

Visualize Free（网址 https://www.visualizefree.com/）是一个免费的可视化工具，其本质上是一个托管平台，允许用户使用公开的或者自行上传的数据集，然后依据设置，构建完成交互式可视化的演示数据。

1.2.3 专业可视化软件

相对于在线可视化工具的单一功能，以下 3 种专业可视化软件的功能则强大得多。

1. Oracle Data Visualization

Oracle Data Visualization 是 Oracle 推出的一款数据可视化独立产品，也是 Oracle BI 产品 BIEE 的一部分。Oracle Data Visualization 的产品组件，不仅仅支持本地部署，也可以在云端方便地访问，甚至在个人计算机的桌面，用户也可以随时随地自如地分析任何来自个人或企业内部的数据。

Oracle Data Visualization 在方便用户使用、加速交互性的同时，可保证数据的准确性和一致性，并具有以下亮点。

可视：让丰富的可视化控件显示数据，并且方便地分享给其他人。

简单：不论是加载数据，或者混搭不同来源的数据，还是以拖动的方式进行交互性探索，都以用户期望的方式进行。

快速：只需要通过点击，就可以快速地检索数据，找到更多的答案和洞察业务。

智能：对数据进行解读，推荐最佳的表现形式，并可以根据上下文自动进行联动。

Oracle Data Visualization 可以有多种部署选择，包括云端的 Data Visualization Cloud Service（DVCS）、本地部署的 Data Visualization（DV）以及桌面版 Data Visualization Desktop（DVD）。用户可以根据自己的实际需要，选择任何一种工作方式，利用相同的技术进行自助式的数据探索，并且可以在不同的工作方式中，非常容易地进行迁移和共享。

2. PowerBI

PowerBI 是微软旗下的一款一体化的 BI 和分析平台，提供“即服务”或者桌面客户端，但是评分最高的还属其可视化功能。可视化能够直接从报告中创建，可以与整个组织的用户共享。

除了大量的内置可视化样式外，也可以在 AppSource 社区不断创建新的可视化样式，或者如果用户想自己编码，那么可以使用开发人员工具（Developer Tools）从头开始创建并与其他用户共享。它还包括一个自然语言界面，允许通过简单的搜索词建立不同复杂度的可视化。

3. Tableau

Tableau 是能够帮助用户查看并理解数据的商业智能软件，具有快速分析、简单易用、不限数据源、智能仪表板、自动更新、瞬时共享等特点。收费版功能较多，有 Tableau Desktop、Tableau Prep、Tableau Online、Tableau Server 等多个版本。

1.3 数据挖掘

1.3.1 数据挖掘概述

在大数据时代，如果人们想要探究数据深层次的内涵，离不开数据挖掘的操作。所谓数据挖掘（Data Mining），又称资料探勘、数据采矿，一般是指从大量的数据中通过算法搜索隐藏于其中的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统和模式识别等方法实现上述目标。

数据挖掘常见的分析方法有：分类、估计、预测、相关性分组或关联规则、聚类复杂数据类型挖掘等。

1.3.2 常用数据挖掘工具

1. IBM SPSS Modeler

IBM SPSS Modeler 是 IBM 开发的一款面向商业用户的高品质数据挖掘工具，该软件拥有可视化用户界面，简单易用，且包含多种挖掘算法，可快速建立数据模型，挖掘结果直观易懂，可应用于商业活动，从而改进决策过程，故在数据挖掘领域具有较高的口碑。

2. R

R 是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言；可操作数据的输入和输出，可实现分支、循环，用户可自定义功能。

3. Oracle Data Mining

Oracle Data Mining 是 Oracle Advanced Analytics 数据库的一个组件，它提供了强大的数据挖掘算法，可以让数据分析师发现洞察、做出预测并利用其 Oracle 数据进行投资。Oracle Data Mining 中的算法以 SQL 函数形式实现，可以挖掘数据表和视图、星状模式数据，包括事务性数据、聚合、非结构化数据以及空间数据。

4. Weka

Weka 是一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则，以及在新的交互式界面上的可视化。Weka 高级用户可以通过 Java 编程和命令行来调用其分析组件。同时，Weka 也为普通用户

提供图形化界面，和 R 相比，Weka 在统计分析方面较弱，但在机器学习方面要强得多。

5. RapidMiner

RapidMiner 是一个用于机器学习和数据挖掘实验的环境，用于研究和实际的数据挖掘任务，是世界领先的数据挖掘开源系统。该工具用 Java 编程语言编写，通过基于模板的框架提供高级分析。

6. KNIME

KNIME 是一个基于 Eclipse 平台开发，模块化的数据挖掘系统，它能够让用户可视化创建数据流，选择性地执行部分或所有分解步骤，然后通过数据和模型上的交互式视图研究执行后的结果。

1.4 数据分析

1.4.1 数据分析概述

所谓数据分析，是指用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总、理解并消化，以求最大化地开发数据的功能，发挥数据的作用。

1.4.2 数据分析的目的与分类

数据分析的目的是把隐藏在大批看似杂乱无章的数据背后的信息集中和提炼出来，总结所研究对象的内在规律，帮助管理者进行有效的判断和决策。

数据分析的分类可分为以下 3 种。

（1）描述性数据分析：侧重于概括和表述数据的整体状况。

（2）探索性数据分析：侧重于在数据中发现新的特征。

（3）验证性数据分析：侧重于验证已有假设的真伪。

1.4.3 数据分析的作用

数据分析的作用主要体现在以下几方面：

1. 市场营销方面

通过数据分析和数据挖掘技术，可以精准寻找目标用户，发现用户特征，构建用户画像，预测用户行为，对用户进行合理分群，用户偏好预测、用户个性化推荐等。

此外，通过对用户行为分析研究，针对用户的多维度属性、标签和行为数据，对用户流失预警、用户生命周期分析、用户影响力分析、用户价值分析等相关用户行为进行研究。

再者，通过监测并分析行业竞品情况，收集并解读相关用户和市场研究报告，为公司产品规划提供支持，对行业竞争品和行情进行监控。

2. 运营管理方面

在运营管理方面，通过对日常报告和数据的制作与维护，运营人员可以对公司业务的运营情况展开深入分析，提出发展策略和建议。借助于监控评估运营活动效能，运营人员也可以评

估运营活动效能，提出营销活动优化和成本控制解决方案，并主导或协助落实。在公司管理层面，通过数据分析，可以针对运营团队整体 KPI 考核及情况制定对应绩效考核方案并跟踪绩效考核实施。

3. 产品研发方面

数据分析可以帮助产品进行优化升级，并对新产品的研发提供有效的数据支持。

4. 大数据平台支持方面

对于基金、证券、期货、投资这些金融行业，每天都会产生大量的数据，这些海量的数据更是离不开数据分析的辅助，对于深层次的数据挖掘具有强大的应用前景。

5. 其他方面

此外，数据分析在餐饮行业、旅游行业、快速消费品行业、教育行业、物流行业、互联网金融行业、建筑业等都具有举足轻重的价值，在如今这个时代，谁先认识到数据分析的巨大潜力并付诸行动，谁就能抢占先机。

1.5 数据分析的步骤

数据分析过程包括 6 个循序渐进的基本步骤，它们缺一不可，相辅相成，也是企业在数据分析时必不可少的步骤。

1. 明确分析目的和思路

明确分析目的和思路有助于帮助分析者提供清晰的指引方向，保证数据分析的有效进行。

2. 数据收集

数据收集是按照确定的数据分析目的收集相关数据的过程，它为数据分析提供基础，一般数据来源于以下 4 个渠道。

（1）权威机构：各国各级政府公开发布的数据，如中国国家统计局等。

（2）互联网：网络平台上公开的数据信息，如微博、百度、大众点评等。

（3）市场调查：自发进行的调研活动，向特定的群体收集数据。

（4）企业数据库：企业掌握的生产、运营数据，一般这类数据不会公开发布，或者，经过脱敏后公开使用。

3. 数据预处理

数据预处理是指对收集到的数据进行加工整理，形成适合数据分析的样式，是数据分析前必不可少的阶段，其目的是从大量的、杂乱无章、难以理解的数据中，抽取并导出对解决问题有价值、有意义的数据，从而提高数据分析的效率。

数据预处理包括数据清洗、数据集成、数据变换和数据归约等。

4. 数据分析

数据分析是指用适当的分析方法及工具，对处理过的数据进行分析，提取有价值的信息，形成有效结论的过程。

数据分析分为以下 3 大类。

（1）描述性数据分析：侧重于概括和表述数据的整体状况，包括数量统计、数据缺失情况、

样本分布、平均值、分位数、方差、指标在时间和空间上的变化趋势等。

（2）探索性数据分析：侧重于在数据中发现新的特征。

（3）验证性数据分析：侧重于验证已有假设的真伪。

5. 数据展现

数据展现在数据分析步骤中是一个重要的角色，只有将收集的数据通过处理和分析，形成有用的信息，并且用图形，如柱形图、饼图、折线图等进行展现，能让人们一目了然地发现数据的本质以及作用，数据展现需要做到内容清晰易理解，信息完整明确、简洁美观。

6. 报告撰写

报告撰写是数据分析的最后一步，是整个数据分析过程的总结，是给企业决策者的一种参考，为决策者提供科学、严谨的决策依据。

一份优秀的数据分析报告，需要有一个明确的主题、一个清晰的目录，图文并茂地阐述数据，条理清晰地呈现结论，使决策者能一目了然地看出报告的核心内容，这样既能给阅读者视觉上的冲击，又能很明确地阐述数据分析的核心内容。最后，需要加上结论以及建议，这样不仅可以给决策者指出问题，还可以提供方案和想法，以便决策者在决策时作为参考。

1.6 数据分析方法论

数据分析方法论是从宏观角度出发，指导数据分析师进行一个完整的数据分析的过程，它是一个指南针，为数据分析师指明数据分析的正确方向。

数据分析方法论是指数据分析的思路，是数据分析的前期规划，指导着后期数据分析工作的开展。数据分析方法论好比装修设计图，它为数据分析工作提供工作框架和指引，而数据分析方法好比装修的工具和技术，它为数据分析提供技术的方法和保障。

1. PEST 分析

PEST 分析是分析企业外部宏观环境的一种方法，虽然不同的企业和行业受宏观环境的影响会有一定的差异，但一般企业和行业进行宏观环境分析时，必然会进行政治环境（Political）、经济环境（Economic）、技术环境（Technological）、社会环境（Social）分析，这四个环境是影响企业的外部环境因素。

2. 5W2H

5W2H 分析法又称七何分析法，是以 5 个 W 开头的英文单词和 2 个 H 开头的英文单词进行提问，从回答中发现问题的线索以及解决方法，它简单、方便、易于理解与使用，广泛用于企业管理和技术活动，对于决策和执行性的活动措施非常有帮助，并且有助于弥补问题的疏漏。

5W2H 指：为什么（Why）、做什么（What）、什么人做（Who）、什么时候（When）、什么地方（Where）、如何做（How）、什么价格（How much）。

3. 逻辑树分析法

逻辑树又称问题树、演绎树或分解树等，逻辑树是将问题的所有子问题分层罗列，从最高层开始，逐步向下扩展，并把一个已知问题当成树干，然后开始考虑这个问题和哪些问题有关，每想到一点，就给这个问题所在的树干加一个“树枝”，并标明这个“树枝”代表什么问题，

一个大的“树枝”上还可以有小的“树枝”，依此类推，找出与问题相关联的所有项目。

逻辑树主要是帮助数据分析师理清自己的思路，避免进行重复和无关的思考。

4. 4P 营销理论

4P 营销理论产生于 20 世纪 60 年代的美国，它是随着营销组合理论的提出而出现的，营销组合实际上有几十个要素，这些要素可以概括为以下 4 类：产品（Product）、价格（Price）、渠道（Place）、宣传（Promotion）。

5. 用户行为理论

用户行为是指用户为获取、使用物品或者服务所采取的各种活动，用户对产品首先需要有一个认知、熟悉的过程，然后试用，再决定是否继续消费使用，最后成为忠诚用户。

1.7 常见数据分析法则

1. 四象限法则

四象限法则是数据分析中经常被用到且非常重要的一个分析方法，在应用上有着多种变化。所谓四象限法则，是指通过对两种维度的划分，运用坐标的方式表达出想要的价值，由价值直接转变为策略，从而进行一些项目的推动。四象限法则是一种策略驱动的思维，广泛应用于战略分析、产品分析、市场分析、客户管理、用户管理、商品管理等，其优点是直观、清晰，可以对数据进行人工划分，划分结果可以直接应用于策略。通过运用四象限法则分析数据，可以快速地找到问题的共性原因，建立分组优化策略。

2. 同期群分析

所谓同期群分析，是指按时间维度对用户建立分组，观察分组用户的行为特征表现，其目的在于透过现象找到结果。以时间维度建立同期群，除按时间维度考虑，也可以按来源渠道等维度建立同期群。

3. 假设分析

在没有直观数据或者线索能进行分析的情况下，可以采用假设分析的方法进行综合考虑，以假设先行的方法进行推断，通过人工设置一个变量来进行反证。例如：新产品的预期销量、未来某段时间内的景区热门度之类的。假设分析是一种启发思考驱动的思维，它更多是一种思考方法，即假设、验证并加以判断。

4. 指数法

指数法主要有线性加权、反比例、log 三种方法，是一种目标驱动的思维，是将无法利用的数据加工成可利用的，从而进行分析。但是指数法没有统一的标准，很多指数更依赖经验来进行加工。指数法的优点是目标驱动力强、直观、简洁、有效，对业务有一定的指导作用，一旦设立指数不易频繁变动。

5. 帕累托法则

帕累托法则，又称二八定律、关键少数法则、不平衡原则等，被广泛应用于社会学及企业管理学等，它以 19 世纪末 20 世纪初意大利经济学家帕累托命名。因为他发现，在任何一组东西中，最重要的只占其中小部分，约 20%，其余 80% 尽管是多数，却是次要的。

帕累托法则是一种只抓重点的思维，应用于绝大多数的领域，因此，这种分析思维几乎没有什么局限性。但是在一些特定的情况下数据分析依旧不能放弃全局，否则就会使思路变得狭隘。

6. 对比分析法

对比分析法是一种挖掘数据规律的思维方式，一次合格的数据分析一般都会用到多次对比，如竞争对手对比、时间同比环比、类别对比、转化对比、特征和属性对比、前后变化的对比等。

在基于相同数据标准下，对比分析由其他影响因素所导致的数据差异，其目的在于找出差异后进一步挖掘差异背后的原因，从而找到优化的方法。

其优点也是显而易见的：对比分析法可以发现很多数据间的规律，可以与任何技巧结合。

7. 漏斗分析

所谓漏斗分析，是一套流程式数据分析，它能够科学反映用户行为状态，以及从起点到终点各阶段用户转化率情况的重要分析模型。漏斗分析模型已经广泛应用于网站用户行为分析和App用户行为分析的流量监控、产品目标转化等日常数据运营与数据分析的工作中。

漏斗分析最常用的是转化率和流失率两个互补型指标。比如有10人访问某电商网站，有3人点击注册，有1人注册成功。这个过程共有三步：第一步到第二步的转化率为30%，流失率为70%；第二步到第三步转化率为33%，流失率为67%；整个过程的转化率为10%，流失率为90%。该模型就是经典的漏斗分析模型。

第2章
数据可视化初步

甲骨文公司，全称甲骨文股份有限公司(甲骨文软件系统有限公司)，是全球最大的企业级软件公司，总部位于美国加利福尼亚州的红木滩。2013年，甲骨文已超越IBM，成为继Microsoft后全球第二大软件公司。

甲骨文公司(Oracle)向一百多个国家的用户提供数据库、工具和应用软件以及相关的咨询、培训和支持服务。

2.1 Oracle DV产品简介

Oracle数据可视化(Oracle Data Visualization)是Oracle公司基于商业智能分析产品BIEE的一个功能扩展。产品于2015年正式发布，是一款集数据整理、数据可视化、数据挖掘(机器学习)为一体的敏捷数据分析软件。

Oracle数据可视化技术有包括云端的Oracle Analytics Cloud(OAC)、本地部署的Data Visualization(DV)以及桌面版Data Visualization Desktop(DVD)多种部署方式。用户可以根据自己的实际需要，选择任何一种部署方式，利用相同的技术进行自助式的数据探索，并且可以在不同的工作方式中，非常容易地进行迁移和共享。

(1)云端部署(Cloud)：Oracle Analytics Cloud(OAC)。

Oracle同时提供标准版、数据湖版、企业版(见图2-1)，除了涵盖DV的可视化和自助分析能力外，还增强了云端的大数据存储及企业级Business Intelligence(BI)分析能力。

(2)桌面版(Desktop)：Data Visualization Desktop(DVD)。

为用户提供了另外一个选择，即在自己的桌面上混搭和分析不同来源的数据，包括个人或部门的数据、企业的数据甚至是来自其他Software-as-a-Service(SaaS)应用的数据。

(3)本地部署(On Premises)：Oracle Data Visualization(DV)。

Oracle Data Visualization(DV)是Oracle Business Intelligence(BI)12c分析平台的一个组成部分，能够帮助用户进一步扩展已经部署的BI平台和前期投资，将数据分析带到一个新的高度。

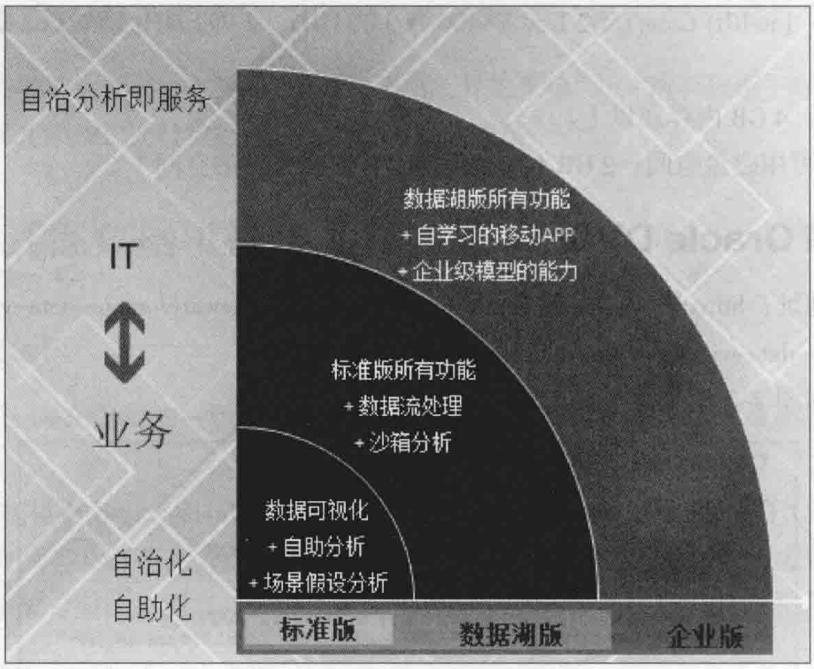

图 2-1 OAC 版本介绍

这里介绍 Oracle Data Visualization Desktop12.2.5.0.0 版本。产品具有以下特征：

（1）可视化：让丰富的可视化控件来讲数据的故事，并且方便地分享给其他人。

（2）简单：不论是加载数据，或者混搭不同来源的数据，还是以拖动的方式进行交互性探索，都以用户期望的方式进行。

（3）快速：只需要通过点击，就可以快速地检索数据，找到更多的答案和业务洞察。

（4）智能：可以智能地对数据进行解读，推荐最佳的表现形式，并可以根据上下文自动进行联动。

相比较于同性质产品，其亮点包括：

（1）可视化图形丰富：提供更丰富、更美观的图形类型，更直观的数据洞察。

（2）贯穿数据分析的全生命周期：提供数据存储、转换、分析及机器学习一站式数据价值获取平台，更便捷的数据价值获取。

（3）机器学习：提供一键式机器学习和算法自定义机器学习平台，让机器学习更简单。

2.2 软件安装

2.2.1 硬件要求

Oracle Data Visualization Desktop 支持 Windows 和 Mac 操作系统下的安装，操作系统及硬件要求如下：

（1）操作系统：Microsoft Windows x64（64 位）7 SP1 +、8.1 或 10；Windows Server 2012 R2；Sierra 10.12、High Sierra 10.13。